

REL Mid-Atlantic Teacher Effectiveness Webinar Series
Using Instructionally Sensitive Assessments to Measure Teacher Effectiveness
Q&A with Dr. W. James Popham
November 13, 2014

This webinar discussed the importance of the instructional sensitivity of assessments and encouraged dialogue about how the use of results from instructionally sensitive assessments can promote teacher effectiveness. This Q&A addressed the questions participants had for Dr. Popham following the webinar. The [webinar recording](#) and [PowerPoint presentation](#) are also available.

Questions

1. How do you establish that an assessment is instructionally sensitive?

There are two processes that establish the validity evidence of an assessment's instructional sensitivity. Instructional sensitivity can be established through judgmental evidence and/or empirical evidence. The process for collecting judgmental evidence involves asking a committee of teachers to indicate item by item if students have been effectively taught to master a standard and if it is likely that a majority of the students could answer the item correctly. The process for collecting empirical evidence involves studying the student scores of two groups of "outlier" teachers (exceptionally successful and exceptionally unsuccessful). If students of teachers in the outlier groups score in unexpected ways on items, a differential item functioning (DIF) statistical analysis allows the detection of instructionally insensitive items.

2. What can an LEA [local education agency] do to make locally developed assessments instructionally sensitive?

An LEA can establish the validity evidence related to an assessment's instructional sensitivity by collecting either judgmental or empirical evidence being used for evaluative purposes. Both of these reviews can dramatically reduce the instructional insensitivity of an assessment. Judgmental evidence is more commonly used for LEAs because empirical evidence requires a large sample size of teachers to compare. To collect judgmental evidence, a committee of teachers could be asked to indicate for each item if students had been taught with skillful pedagogy to master the content standard assessed by the item. Teachers respond by indicating whether a substantial majority of students could answer the question successfully.

3. Are there templates or models for creating instructionally sensitive assessments?

There are no specific templates or models to create instructionally sensitive assessments, as assessments vary by content area and grade, among other factors. There are review processes that assessments can go through to increase the likelihood of instructional sensitivity.

4. How do we ensure cultural sensitivity on assessments?

Through the reviews involved in collecting validity evidence (judgmental or empirical), cultural sensitivity can be tested as well. Teachers can be directed to review items for

cultural sensitivity.

5. How are assessments used to improve teachers' awareness of their instructional impact?

Instructional sensitivity is the degree to which students' performances on a test reflect the quality of the instruction specifically provided to promote students' mastery of what is being assessed (Popham, 2007). Assessment data can provide teachers with information about their instructional impact because they know what they have taught their students and they can see the effects of their instruction based on assessment results (assuming the assessments have been tested for instructional sensitivity).

6. How can data help teachers set learning targets?

Instructionally sensitive assessments can provide teachers with information regarding their pedagogy and what standards they are successfully teaching. Items that their students miss can indicate either a weakness in instruction or a missing piece of instruction (assuming the assessment is instructionally sensitive), and learning targets can be based on an analysis of these data. If a test has been made instructionally sensitive, and teachers still are unable to get many students to master a given learning target, it is possible that the teacher has chosen an inappropriate learning target.

7. How can we discuss assessments with students and/or parents?

The National Assessment Governing Board (<https://www.nagb.org/>) intends to provide assessment-related resources and information for parents and students to use in discussing assessment in students' families. The project intended to supply these assessment-literacy resources is just getting underway; hence, the resources will not be available in the immediate future.

8. How can generalized norm-referenced student test scores be integrated in teacher evaluation systems? How can other data sources be included with high validity in teacher evaluation?

When using assessment scores in teacher evaluation systems, it is crucial to ensure that the assessments used are instructionally sensitive and that validity evidence supports the claims made. The book *Evaluating America's Teachers: Mission Possible?* discusses important considerations related to validity in teacher evaluation. Assessments are sometimes used to evaluate teachers when the validity of the assessment has not been established. Not all assessments are valid for evaluating teachers and schools. To use assessments in teacher or school evaluation, the assessments should be tested for validity for an evaluative purpose. This typically translates into the need for instructional sensitivity evidence.

9. Should tests be sensitive or insensitive to instruction when used for accountability purposes?

Tests should be sensitive to instruction because when tests are sensitive to instruction, the difference between well taught and poorly taught students can be revealed and understood. Tests that are insensitive to instruction will mask, not illuminate the caliber of instruction.

Insensitive tests must not be used to evaluate teachers or schools. Yet, such an appalling misuse of educational testing is rampant in the United States.

10. Do student evaluations and surveys of teachers have high validity?

Student surveys can contribute to valid inferences about teachers' skill when care is taken in developing the types of surveys and the directions for their uses. The surveys should be anonymous, accessible to students, and presented to students in a way they can understand. Students will, early on, make mistakes because they are not accustomed to evaluating their teachers, but too-high and too-low results can average out to provide an accurate picture of the teacher's instruction.

11. How can we ensure that money and resources are effectively used in assessment development?

The key question in test development is the quality of the product. All stakeholders need to be smarter about what they demand from testing companies. Both sides must be clear about what is being developed and for what test-based inferences and actions. Tests should be developed for a certain purpose, and alignment and validity evidence should be on hand so that the test is not used for inappropriate purposes. Reviewing for instructional sensitivity does not cost much during assessment development. When assessments are not tested for instructional sensitivity, it is often because of the lack of knowledge and zeal around instructional sensitivity.

12. Are there any particularly applicable sections in the new 2014 standards?

The validity chapter in the new *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) is especially useful and important.

13. What do you think of value-added measures?

Value-added measures have their limitations and make it difficult to measure a teacher in a specific setting. When value-added measures are used, they should be used with instructionally sensitive assessments. Just as we can't make chocolate-fudge brownies out of a lump of Silly Putty, we can't employ value-added statistical analyses to make evaluative sense out of students' performances on instructionally insensitive tests.

14. In what ways are these reviews different from teacher committee item reviews and benchmarking that occur and have occurred in most states?

Collecting judgmental evidence is not different from reviewing items as we now routinely do. In the collection of judgmental evidence, reviewers are oriented in what they should be checking. They check for items that might test the student's IQ or background as opposed to the student's learning. Reviewers are given a rubric, checklist, and/or training to indicate what types of items they are looking for. We have relatively little experience in knowing how to best undertake judgmental analyses of test items' instructional sensitivity.

15. What can stakeholders do to ensure that assessments are instructionally sensitive?

Stakeholders can demand instructional sensitivity evidence before sanctioning the evaluative use of a test. Stakeholders can learn about the issue of instructional sensitivity and insist on this crucial validity evidence before using assessments.

Action Steps

Participants responded to the question “As a result of today’s webinar, what action steps do you plan to take?” and some of their responses are listed below.

- *Participate in assessment development.*
- *Be a more critical consumer of tests and ensuring that tests are being used for purposes for which there is evidence to support them.*
- *Keep informed with what my state is planning to do.*
- *Share this information with my administrator and my co-teachers.*
- *Get my hands on the revised Standards for Educational and Psychological Testing and get reading.*
- *Take on a better role in evaluating judgmental evidence.*
- *Review this presentation, complete further reading, and think about how this is being put in place and how I can help teachers understand it.*
- *Incorporate these gems in courses/experiences for pre-service teachers in teacher education programs (e.g., assessment course/experiences for teachers).*

Additional Resources

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, 89, 146–155.
- Popham, J. W. (2013). *Evaluating America’s teachers: Mission possible?* Thousand Oaks, CA: Corwin.